Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

From Lurkers to Workers: Predicting Voluntary Contribution and Community Welfare

Marios Kokkodis, Theodoros Lappas, Sam Ransbotham

To cite this article:
Marios Kokkodis, Theodoros Lappas, Sam Ransbotham (2020) From Lurkers to Workers: Predicting Voluntary Contribution and Community Welfare. Information Systems Research 31(2):607-626. https://doi.org/10.1287/isre.2019.0905

informs
http://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Vol. 31, No. 2, June 2020, pp. 607–626
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

From Lurkers to Workers: Predicting Voluntary Contribution and Community Welfare

Marios Kokkodis,[a] Theodoros Lappas,[b] Sam Ransbotham[a]

[a] Carroll School of Management, Boston College, Chestnut Hill, Massachusetts 02467; [b] School of Business, Stevens Institute of Technology, Hoboken, New Jersey 07030

Contact: kokkodis@bc.edu, https://orcid.org/0000-0002-5037-6060 (MK); tlappas@stevens.edu, https://orcid.org/0000-0002-4669-4170 (TL); sam.ransbotham@bc.edu, https://orcid.org/0000-0001-5305-035X (SR)

Received: April 5, 2017
Revised: March 3, 2018; January 14, 2019; September 9, 2019
Accepted: September 25, 2019
Published Online in Articles in Advance: June 4, 2020

https://doi.org/10.1287/isre.2019.0905

Abstract. In an online community, users can interact with fellow community members by voluntarily contributing to existing discussion threads or by starting new threads. In practice, however, the vast majority of a community's users (≈90%) remain inactive (lurk), simply observing contributions made by intermittent (≈9%) and heavy (≈1%) contributors. Our research examines increases and decreases of types of user engagement in online communities using hidden Markov models. These models characterize latent states of user engagement from trace user activity or lack of activity. The resulting framework then differentiates lurkers who can later become workers (i.e., engaged in the community) from those who will not. Differentiating lurkers who can be engaged from those who cannot enables managers to anticipate and proactively direct their resources toward the users who are most likely to become or remain workers (i.e., heavy contributors), thereby promoting community welfare. Analysis of 533,714 posts from an online diabetes community shows that incorporating latent user engagement variables can significantly improve the accuracy of welfare prediction models and guide managerial interventions. Application of our framework to five additional communities of various contexts demonstrates its generalizability.

History: Alok Gupta, Senior Editor; Maytal Saar-Tsechansky, Associate Editor.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2019.0905.

Keywords: online communities • welfare of online communities • voluntary online work • predictive modeling

1. Introduction

Millions of users engage daily with online communities through forums that span an ever-increasing variety of topics (Wikipedia 2018). These communities rely on their users to generate forum content through voluntarily working by sharing, responding, answering, and discussing topics of interest. These activities engage users as they interact with each other, ideally leading to cohesive and productive communities (Leimeister et al. 2006, Lin and Lee 2006, Zhang and Watts 2008, Hew 2009, Seraj 2012).

The emergence of these online communities challenged existing theories about community participation (Faraj et al. 2011). The proliferation of information and communication technologies both dramatically- increased the reach of communities and simultaneously reduced costs of contributing to them. For example, participating in a community with globally dispersed members no longer requires travel to synchronously meet. Instead, people can use general purpose technologies at a time of their convenience from wherever they are.

However, technology did not eliminate all costs. Although the mechanical aspects of contributing are cheaper, there may still be personal costs of contributing. For example, people may be "too shy to contribute" (Sun et al. 2014) as the global reach of the community increases social exposure of contributors. Indeed, despite the reduced costs of contributions, the vast majority of users consume content without contributing (Van Mierlo 2014). Pervasiveness of this imbalance gave rise to an aphorism, the 1-9-90 principle (Wu 2018), which postulates that 1% of an online community's user base generates original content, an additional 9% only interacts with existing content introduced by others, and the remaining 90% of the participants passively lurk. As a result, the global reach of communities is not uniformly positive because users, although benefiting from the community presence, may be reluctant to contribute their own work.

Research has made considerable progress decrypting the process of converting these passive users into engaged contributors. Theoretical models describe engagement levels in numerous ways, such as "reader to leader" (Preece and Shneiderman 2009), core periphery (Lave and Wenger 1991), and so forth. Important topics focus on the determinants (Wasko and Faraj 2005, Moon and Sproull 2008, Bateman et al. 2011) and motivations (Wang and Fesenmaier 2003) of user engagement as well as the emergence of leadership (Lu et al. 2013, Johnson et al. 2015).

Despite these findings, previous research focuses primarily on participation and observable user actions and has only recently begun to model lurkers explicitly (Tagarelli and Interdonato 2014). However, focusing on observable actions misses many potential users, whereas such observable actions can be incidental activity bursts that may not necessarily reflect a user's underlying dynamic engagement with the community (Chen et al. 2018). Furthermore, much of the prior research considers all types of contribution equally (e.g., "number of responses"). This creates the opportunity to better understand community participation (Malinen 2015) because personal costs vary considerably by contribution type. Finally, even though previous work extensively discusses community welfare measures (Preece 2001, Szmigin et al. 2005, Wise et al. 2006, Chai et al. 2011), a formal framework that predicts welfare could guide managerial intervention.

To address these limitations, we propose a dynamic framework that models both observed participatory actions and latent engagement states for users, including lurkers. We motivate and design a two-layer hidden Markov model (HMM) customized for the online community context. The HMM allows transitions between latent states that we describe through accelerated failure time (AFT) models. Extending previous research that considers one type of participation (Chen et al. 2018), topic models cluster contribution into four groups: "Ask," "Share," "Respond," and "Append." The resulting HMM-AFT framework accurately predicts user transitions among latent engagement states based on their type of contribution.

Additionally, the ability to model engagement at the user level has implications for the study of a community as a whole. Community welfare can be measured in many ways, such as via contribution volume at the community level (Wiertz and de Ruyter 2007), user interactivity (Preece 2001, Wise et al. 2006), and reciprocity (Chai et al. 2011, Faraj and Johnson 2011). Despite their differences, these definitions are consistently based on user contribution and hence, user engagement. Therefore, we expect that a community's welfare depends on the presence of highly engaged users. This motivates us to use user engagement as estimated by our framework toward early prediction of community welfare. To model user engagement and its predictive properties, we focus on an important health-related context—DiabetesForum (pseudonym), an online community dedicated to diabetes. Analysis of 533,714 posts from this community shows that incorporating user engagement variables can significantly improve the accuracy of welfare prediction models and successfully guide managerial interventions. Application of the HMM-AFT framework in five additional communities of various contexts demonstrates its generalizability.

The results add to our understanding of online communities. First, by explicitly modeling lurkers through a dynamic framework of latent user engagement, we differentiate lurkers who can be engaged from those who cannot. Second, we differentiate types of contribution by the confidence that they require, thereby allowing for targeted intervention. Third, we dynamically model latent patterns of community activity to create a framework for predicting and evaluating community welfare. Fourth, the proposed design guides practitioners to efficiently address challenges when modeling user participation and engagement in online communities.

2. Theoretical Context

2.1. Benefits and Costs of Online Communities to Users

To the benefit of many users around the world, the proliferation of internet access brought a concomitant rise in online communities organized around many topics. This rise motivated researchers to investigate the many new and diverse phenomena that prior offline theory could not explain well. Examples include investigations of user willingness to pay (Oestreicher-Singer and Zalmanson 2013), foundational questions of defining user engagement (Ray et al. 2014), comparisons of work-related versus leisure-related content creation (Huang et al. 2015), sharing content between similar and dissimilar users in terms of expertise (Hwang et al. 2015), contrasts in mobile versus nonmobile content generation and usage (Ghose and Han 2011, Ransbotham et al. 2019), and even distinctions in content generation as virtual communities distance themselves from the physical world (Kohler et al. 2011).

However, although information and communication technologies both dramatically increased the reach of communities and simultaneously reduced costs of contributing to them, costs have not been eliminated. Although the mechanical aspects of contributing are cheaper, people still are reluctant to contribute because of personal costs (Sun et al. 2014). Thus, the global reach of communities is not uniformly positive because users, although benefiting from the reach of the community, may be reluctant to contribute their own work exactly because of the global exposure (Nonnecke and Preece 2001). In fact, the pervasiveness of computing (Ransbotham et al. 2016) leads to vulnerability risks (Cramer and Hayes 2010), such as cyberbullying (Willard 2007, Hay et al. 2010), personal insults, and other threats (Patton et al. 2013, Isaacs 2014). These potential personal costs affect the willingness of users to engage in communities, which in turn, affects the communities' welfare.

2.2. Engaging Users to Contribute

Given these costs, two challenges that online communities face are (1) motivating users to start contributing and (2) engaging users after they contribute so that they remain engaged. User engagement is thus a continuous and uncertain effort; even when communities have sufficient resources to regularly engage their users, only a small fraction eventually become and remain engaged. The majority of users lurk, observing community activity but not contributing (Van Mierlo 2014).

As a result, an ongoing and rich literature explores the mechanisms that motivate users to contribute their voluntary work to a community. For example, contribution can result in social capital benefits; users may contribute knowledge out of a desire to grow their professional reputation (Wasko and Faraj 2005). Alternatively, users may be driven by psychological and interpersonal bonds that they forge when they identify with various users or groups within the community (Bateman et al. 2011, Ren et al. 2012). This identification may develop through the community itself, or it might preexist through social ties formed in offline contexts (Bagozzi and Dholakia 2006, Zeng and Wei 2013). Furthermore, because individuals are heterogeneous, their idiosyncratic cognitive, emotional, and social characteristics affect their decisions to contribute (Bagozzi and Dholakia 2006, Tsai and Bagozzi 2014). However, an important difficulty is that, in many cases, the underlying causes of participation owing to intrinsic motivation and interest are not observable externally or measurable by platform managers. Our analysis (Section 3.1) acknowledges the latent aspect of these user attributes because it models user evolution from the initial decision to engage with the community through subsequent decisions that define their dynamic level of engagement.

Even though idiosyncratic reasons may drive motivation to participate in online communities, community characteristics also influence users' decision to engage. Social learning theory indicates that users can learn about the behavior of other users even by lurking. These lurkers can absorb information as they lurk. For example, community feedback is particularly important to user engagement in diverse contexts, such as initial (Lampe and Johnston 2005) and subsequent (Joyce and Kraut 2006) contribution to news forums, contributions to Wikipedia (Ozturk and Nickerson 2015), contribution to technical support forums (Moon and Sproull 2008), and even participation in digital workplaces (Kokkodis and Ipeirotis 2016, Kokkodis and Ransbotham 2019). This feedback affects both the quantity and quality of contributions (Arguello et al. 2006, Moon and Sproull 2008, Burke et al. 2009, Huberman et al. 2009). From a social learning theory perspective, feedback reinforces behavior. Importantly, users who have not yet contributed can observe these community characteristics. Lurkers who observe the feedback of others learn the activities that lead to positive reinforcement. More engaged users receive the feedback directly and learn which behaviors the community rewards. The proposed framework (Section 4.2) models observed and reinforced behaviors on user engagement: community attributes that measure the reinforcement improve the predictive accuracy of a user engagement dynamic model.

2.3. Levels of Engagement

Additionally, user engagement is unlikely to be binary with users engaging either as active contributors or not. Instead, engagement likely progresses through multiple states of increasing activity. For example, users become more competent as they become more engaged in the main processes of the particular community. They move from legitimate peripheral participation to full participation (Lave and Wenger 1991). Subsequent analysis, for example, differentiates between visitor, novice, regular, and leader roles (Kim 2000) and develops a reader-to-leader framework with emphasis on different needs and values at different participation levels (Preece and Shneiderman 2009).

From this perspective, the heaviest contributors are arguably a community's most valuable demographic because (by definition) they generate the vast majority of the community's content (Yoo and Alavi 2004, Cassell et al. 2006). Therefore, understanding the emergence of these contributors as well as their patterns of participation is important. Contributors emerge because of their social capital, sociability, and knowledge contributions (Faraj et al. 2015). These heavy contributors affect the community both directly and indirectly.

Direct contributions by heavy contributors are certainly foundational. When adding content, heavy contributors tend to use multiple discourse channels to broadcast their contributions (Forte and Bruckman 2005), use language familiar to the rest of the community (Johnson et al. 2015), and tend to submit more and better content (Goes et al. 2014).

However, beyond their direct contribution, heavy contributors also provide positive feedback that generates local network effects in content generation through social learning (Shriver et al. 2013). These network effects build on explicit and implicit ties that users form between each other (Adler and Kwon 2002, Reagans and McEvily 2003, Grewal et al. 2006). Explicitly, many platforms allow users to follow each other; a stochastic network growth model, for example, can predict the formation of these ties (e.g., in an online reviewing platform) (Lu et al. 2013). Implicitly, tacit knowledge of processes and community practices transfers between artifacts of user-generated content (e.g., between Wikipedia articles) (Ransbotham et al. 2012). Because users absorb information through both explicit and implicit network mechanisms, each of these mechanisms can affect the resulting user-generated content indirectly.

2.4. Modeling Lurkers

Despite the large amount of research conducted on user engagement, "a conceptual framework for user participation remains undefined as most of the research has approached participation in terms of its quantity" (Malinen 2015). With a focus on quantity and heavy contributors, lurkers receive less attention, perhaps exactly because quantity of activity is easier to measure than the absence of activity. Exacerbating this difficulty, the term "contribution" often groups different types of participation (e.g., questions, shares, responses). Because lurkers differ significantly from contributors in terms of their willingness to share information and their motivation to join the community (Ridings et al. 2006, Sun et al. 2014, Phang et al. 2015), there is a need for a framework that can model engagement in the absence of participation.

Our framework models engagement as a latent state. Quantities of activities are important signals, but they are imperfect. Other research is beginning to recognize the potential value from a latent approach. For example, a dynamic approach across three latent motivation states in a question-and-answer forum finds that reciprocity and peer recognition influence user motivation to answer questions (Chen et al. 2018). We extend this approach in several ways. First, instead of considering only one type of contribution (i.e., responses), we consider different types of contribution, such as "Ask," "Share," "Respond," and so forth. Second, we explicitly model lurkers (rather than only contributors). Modeling lurkers is crucial from a managerial perspective, because it provides a realistic evaluation of the community status.

Importantly, our proposed framework stochastically separates lurkers who are unlikely to ever become contributors from those who are eventually engaged. This distinction can be critical for efficient interventions (Section 5.4). In addition, the proposed two-layered structure reduces noise in estimating the initial intentions of each user, which increases prediction accuracy. Finally, our research considers a series of alternative modeling choices (e.g., choice of the survival distribution) and an extended set of observed factors that motivate transitions to new states. Evaluation in six different communities shows that the proposed framework outperforms prior frameworks in both predicting individual engagement and measuring a community's welfare.

Finally, lurkers exist in all kinds of online communities. In social networks, for example, graphs describe passive connection between "friends" (e.g., Facebook) or "followers" (e.g., Twitter). Relevant research in these networks focuses on ranking lurkers according to their network connections, similar to how the Pagerank algorithm ranks web pages (Tagarelli and Interdonato 2013, 2014, 2015). Delurking in these networks can be done computationally (Interdonato et al. 2015) or through users outside the focal network (i.e., the friends of a user) (Interdonato et al. 2016). Application of these approaches to question-answering communities (which are the focus of this research) is not straightforward because one needs to define the structure of a network. Because friendships are not observed in online forums, one way to create network connections is by connecting "askers" with "responders." This structure ignores lurkers, who end up having no connections. Hence, such network approaches fail to perform well in predicting user engagement in online question-and-answer communities (Section 5).

2.5. Online Community Welfare

The extensive body of work on behavior and contribution at the user level is complemented by studies that focus on the welfare of the community as a whole. Community welfare is a dynamic concept that varies as the community evolves. It is thus measured in the context of a specific timeframe. From a managerial perspective, monitoring a community's welfare and intervening accordingly are important (Beenen et al. 2006, Seraj 2012, Healey et al. 2014). As soon as managers become aware of a community's deteriorating welfare, they can take remedial actions, such as the introduction of new features or incentives (Cheng and Vassileva 2006, Janzik and Herstatt 2008). Thus, *predicting* the community's welfare given its current status is managerially important.

The ability to monitor and predict a community's welfare assumes a formal operationalization of the concept. Prior research proposes various alternatives for measuring community welfare, such as user activity (Preece 2001), interactivity (Szmigin et al. 2005, Wise et al. 2006), and reciprocity (Wiertz and de Ruyter 2007, Chai et al. 2011, Faraj and Johnson 2011). The number of new threads and the number of responses that users contribute during a given timeframe (Ridings et al. 2006, Saltz et al. 2007, Shen and Khalifa 2007, Himelboim et al. 2009, Millington 2012, Schneider et al. 2013) often measure user activity.

However, posts that users contribute either as a response to the creator of a thread or as a response to a previous responder (Preece 2001, Viégas and Smith 2004, Angeletou et al. 2011, Bernstein et al. 2011, Cheng et al. 2015) measure interactivity. Finally, reciprocal actions between users (e.g., a user x responds to a post by user y after y had responded to a post by x) (Hemetsberger et al. 2002, Lampel and Bhalla 2007, Preece and Shneiderman 2009, Pai and Tsai 2016) measure reciprocity.

Even though each measure describes a different aspect of community welfare, they are all based on the users' contributions patterns: a community with highly engaged, active users is likely to record high scores for all of these measures. Even further, a community that has consistently maintained a high number of highly engaged users is more likely to achieve and maintain high welfare levels in the future. Our framework allows us to accurately estimate the number of highly engaged users at any point and significantly improves community welfare prediction (Section 5.3.2).

2.6. Summary

Although the studies mentioned in this section examine user engagement in online communities, each differs in emphasis and perspective. Table 1 summarizes these studies with five notable dimensions. The first dimension ("lurker modeling") indicates if the study examines lurkers explicitly. The second ("user evolution") specifies whether the research explicitly models the dynamic user behavior in online communities. The third ("which users") and fourth ("when active") dimensions identify whether the research predicts which users and when will become more engaged. The fifth dimension ("welfare") indicates whether the research focuses on predicting community welfare attributes. Finally, the last dimension ("contribution types") indicates whether the research studies engagement in relation to different types of contribution. Taken together, they illustrate the many aspects of online communities and the variety of perspectives through which researchers can study these aspects.

3. Methodology

User engagement is a latent (unobserved) dynamic variable that, rather than being directly measurable, can be inferred from the observed user activity. Our research models this latent engagement through an HMM built from trace data to examine how user engagement relates to dimensions of community welfare (activity, interactivity, and reciprocity).

3.1. Modeling User Engagement

Users join an online community with different objectives; some users join to ask questions, others join to find information in the community's existing threads, and still others join to share their knowledge with those asking questions (Malinen 2015). These intrinsic motives and objectives are both inherently *unobserved* (i.e., not directly stated or even recognized by the users themselves) and *dynamic* (i.e., evolve with time). For example, users who join the community to respond to a specific question might lose interest, or they might start responding to new questions. Similarly, users who join the community to passively learn about a topic might later feel confident enough to respond to questions. Even though we cannot observe the actual state of each user, we observe the trace of user activity that reveals information about their state of engagement with the community.

3.1.1. A Hidden Markov Model of User Engagement. An HMM can help understand community user engagement. HMMs fit this context well because they formally capture dynamic transitions of users over a set of unobserved latent states through a series of observed signals. An HMM assumes a set of states $\mathcal{S} = \{s_1, \ldots, s_K\}$. At any given point in time t, a community user operates from an unobserved engagement state $S_t \in \mathcal{S}$. Each state represents different probability distributions over a set of observable actions, $Y_t \in \mathcal{Y}$.

In an online community, (1) a user might create one or more new discussion threads by either asking a question ("Ask") or sharing information ("Share"). (2) A user might respond first ("Respond") to one or more existing discussion threads. (3) A user might contribute to an ongoing conversation that already has a response ("Append"). (4) A user might decide to be passive ("Lurk"). These signals link well to the confidence that a user has in participating in the online community (Section 2.1). For example, lurking requires little self-confidence because users can anonymously consume community content. Responding to a thread reflects increasing amounts of confidence. A user needs more confidence to be the 1st responder compared with being the 10th or 20th responder in a thread. Initiating a thread (by either asking a question or sharing information) reflects even greater confidence. Hence, a potential set of observable actions (or lack of actions) that captures increasing levels of confidence is

$$\mathcal{Y} = \{\text{Lurk}, \text{Append}, \text{Respond}, \text{Ask}, \text{Share}\}. \quad (1)$$

3.1.2. Model Structure. Each user who joins the community has unobserved latent objectives. As the user spends time on the platform, we observe signals of these objectives. We encode this behavior in a two-layer structure (as in other environments) (Kokkodis 2018, 2019b). The first layer is an initial latent state s_1 for all

Table 1. Focal Literature on Online Communities

Paper	Data	Lurker modeling	User evolution	Which users	When active	Welfare	Contribution types	Objective	Methodology
Lu et al. (2013)	E-pinions (online reviews)	✗	✓	✓	✓	✗	✗	Leader emergence	Net growth PSA
Wasko and Faraj (2005)	Private community (legal advice)	✗	✗	✗	✗	✗	✗	Why users contribute	PLS
Bateman et al. (2011)	Survey on Q&A users (n = 324)	✗	✗	✗	✗	✗	✗	Why users contribute	PLS
Zeng and Wei (2013)	Flickr (photos)	✗	✗	✓	✓	✗	✗	Social ties and content	FE panel
Tsai and Bagozzi (2014)	Survey on VC users (n = 982)	✗	✗	✗	✗	✗	✗	Why users contribute	SEM
Ray et al. (2014)	Survey on WoM and Q&A users (n = 301)	✗	✗	✗	✗	✗	✗	Understand engagement	SEM
Oestreicher-Singer and Zalmanson (2013)	Last FM (social network)	✗	✗	✗	✓	✗	✗	Willingness to pay and participation	Logit, PSM, Cox
Preece and Shneiderman (2009)	No data	✗	✗	✗	✗	✗	✗	Understand participation	NA
Bagozzi and Dholakia (2006)	Survey on Linux users (survey, n = 401)	✗	✗	✗	✗	✗	✗	Understand participation	SEM
Moon and Sproull (2008)	Www, Lsoft, Tech, Comp (technical support Q&A)	✗	✗	✓	✗	✗	✗	Feedback and UGC	Cox model
Burtch et al. (2017)	Retail platform (online reviews)	✗	✗	✗	✗	✗	✗	Social and monetary incentives	RE
Ren et al. (2012)	MovieLens (movie-related community)	✗	✗	✗	✗	✗	✗	Member attachment	RE
Goes et al. (2014)	E-pinions (online reviews)	✗	✗	✗	✗	✗	✗	User popularity and reviews	Panel, matching
Shriver et al. (2013)	Soulrider (sports community)	✗	✗	✗	✗	✗	✗	Social ties and UGC	Regression, instruments
Chen et al. (2018)	SuperUser (computer Q&A)	✗	✓	✓	✓	✗	✗	Model active users' motivation states	HMM

Notes. The column "Lurker modeling" identifies whether the research explicitly models lurkers. The column "User evolution" captures whether the research studies the dynamic behavior of users. Columns "Which" and "When" refer to whether the research identifies and predicts which users and when they will become highly engaged. The "Welfare" column shows whether the research focuses on measuring the welfare dimensions of a community. Finally, the "Contribution types" column shows whether the research studies engagement in relation to different contribution types. FE, fixed effects; PSA, parametric survival analysis; Q&A, question-and-answer forum; RE, randomized experiment; PLS, partial least squares; WoM, word of mouth; SEM, simultaneous equations model; PSM, propensity score matching; UGC, user generated context VC, virtual community; NA, non-applicable; ✗, main model focuses only on contributors, but it can generalize on lurkers.

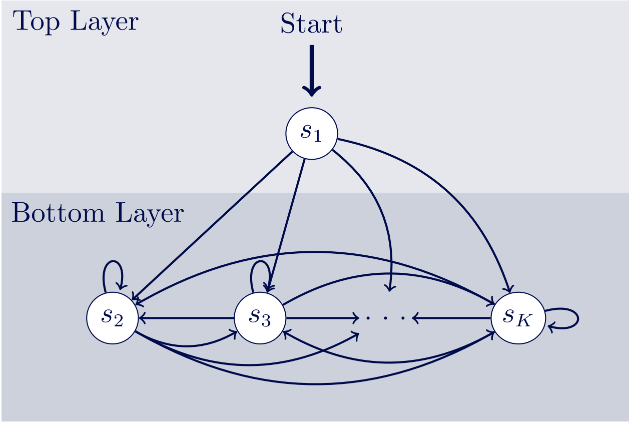
new users. This state allows for an initial estimate of a user's latent objectives. (Alternatively, users could begin stochastically in a second-layer state; this would increase noise in subsequent predictions without adding information.) After a time t, users emit their first signal $Y_1 \in \mathcal{Y}$ and then, stochastically transition to one of the $K-1$ states of the second layer. Figure 1 illustrates this by distinguishing the two layers and depicting the possible transitions from each state.

A complete definition of an HMM requires (1) a vector of initial state probabilities π, (2) a transition matrix T of the transition probabilities between states, and (3) an emission matrix E that describes the state-specific probability distributions across the set of actions $\mathcal{Y}$. Because every new user begins in state s_1, the initial probability vector for the HMM is $\pi = [1, 0, 0, \ldots, 0]$.

A community user's history provides multiple observable signals that correlate with transitions to higher (or lower) engagement states (e.g., time from last action, total number of actions, distance between actions, etc.). Such historical attributes define a vector $\boldsymbol{X}_t$, which affects transition probabilities through, for example, an AFT model (Mario et al. 2008).

AFT models assume that the survival probability of a user follows a specific distribution f, with a cumulative distribution F. For a user in state s_k, time

Figure 1. (Color online) A Two-Layer Hidden Markov Model of User Engagement in an Online Community



Notes. The top layer consists of a single initial state s_1, which represents the starting point for all new community users. After the users emit their first observable action $Y_1 \in \mathcal{Y}$, they transition to an appropriate state in the bottom layer. After this transition, users do not return to the initial state but instead, stochastically transition among the other $K - 1$ available states.

accelerates or decelerates depending on vector $\boldsymbol{X}_t$ (Mario et al. 2008):

$$\tau_{s_k} = \exp\left(-\boldsymbol{\beta}_{s_k}\boldsymbol{X}_t\right)t, \tag{2}$$

where $\tau_{s_k} \sim f$. The survival probability of this user is then given by the following:

$$Surv(t|\boldsymbol{X}_t) = 1 - F\left(\exp\left(-\boldsymbol{\beta}_{s_k}\boldsymbol{X}_t\right)t\right). \tag{3}$$

Assuming an ordering of states in terms of contribution, the probability of transitioning to a state l (i.e., not surviving) with higher contribution is

$$\begin{aligned} &\lambda^{s_k,s_l}_{\boldsymbol{\Theta}\boldsymbol{X}_t} := \Pr(S_{t+1} = s_l | S_t = s_k; \boldsymbol{\Theta}, \boldsymbol{X}_t) \\ \Leftrightarrow\ &\lambda^{s_k,s_l}_{\boldsymbol{\Theta}\boldsymbol{X}_t} = \Pr\left(\xi_{l-1} < \exp\left(-\boldsymbol{\beta}_{s_k}\boldsymbol{X}_t\right)t < \xi_l\right) \\ &= \Pr\left(\exp\left(-\boldsymbol{\beta}_{s_k}\boldsymbol{X}_t\right)t < \xi_l\right) \\ &\quad - \Pr\left(\exp\left(-\boldsymbol{\beta}_{s_k}\boldsymbol{X}_t\right)t < \xi_{l-1}\right) \\ &= F\left(\xi_l - \exp\left(-\boldsymbol{\beta}_{s_k}\boldsymbol{X}_t\right)t\right) \\ &\quad - F\left(\xi_{l-1} - \exp\left(-\boldsymbol{\beta}_{s_k}\boldsymbol{X}_t\right)t\right), \end{aligned} \tag{4}$$

where ξ_l are positive ordered thresholds, such as $\xi_l > \xi_{l-1} \forall l \in \{1, \ldots, K-1\}$ ($\xi_0 = 0, \xi_K = \infty$). For notation simplicity, we group all of the parameters in vector $\Theta = [\boldsymbol{\beta}_{s_1}, \boldsymbol{\beta}_{s_2}, \ldots, \boldsymbol{\beta}_{s_K}, \xi_1, \xi_2, \ldots, \xi_{K-1}]'$.

Figure 2 shows the HMM-AFT framework that includes the structural interactions of covariates $\boldsymbol{X}_t$ with the transition probabilities to different states. Because of the two-layer structure (Figure 1), the transition matrix is not completely filled; instead, the transition matrix is

$$T(\boldsymbol{\Theta}, \boldsymbol{X}_{t-1}) = \begin{bmatrix} 0 & \lambda^{s_1,s_2}_{\boldsymbol{\Theta}\boldsymbol{X}_{t-1}} & \cdots & \lambda^{s_1,s_K}_{\boldsymbol{\Theta}\boldsymbol{X}_{t-1}} \\ 0 & \lambda^{s_2,s_2}_{\boldsymbol{\Theta}\boldsymbol{X}_{t-1}} & \cdots & \lambda^{s_2,s_K}_{\boldsymbol{\Theta}\boldsymbol{X}_{t-1}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \lambda^{s_K,s_2}_{\boldsymbol{\Theta}\boldsymbol{X}_{t-1}} & \cdots & \lambda^{s_K,s_K}_{\boldsymbol{X}_{t-1}} \end{bmatrix}. \tag{5}$$

The emission matrix consists of elements with the conditional probabilities of actions given the current state of the user. These emission probabilities follow a multinomial distribution across the set of available actions $\mathcal{Y}$. Formally,

$$E(\boldsymbol{\mu}) = \begin{bmatrix} \mu^{s_1}_{\text{Lurk}} & \mu^{s_1}_{\text{Ask}} & \mu^{s_1}_{\text{Share}} & \mu^{s_1}_{\text{Respond}} & \mu^{s_1}_{\text{Append}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu^{s_K}_{\text{Lurk}} & \mu^{s_K}_{\text{Ask}} & \mu^{s_K}_{\text{Share}} & \mu^{s_K}_{\text{Respond}} & \mu^{s_K}_{\text{Append}} \end{bmatrix}, \tag{6}$$

where the emission probability of action y at state s_k is $\mu^{s_k}_y = \Pr(Y = y | S = s_k)$, $y \in \mathcal{Y}$, $s_k \in \mathcal{S}$, and $\boldsymbol{\mu} = [\mu^{s_1}_{\text{Lurk}}, \mu^{s_1}_{\text{Ask}}, \mu^{s_1}_{\text{Share}}, \ldots, \mu^{s_K}_{\text{Append}}]$.

3.1.3. Model Identification.

To estimate the parameter vectors $\boldsymbol{\Theta}$ and $\boldsymbol{\mu}$, we maximize the conditional probability of the observed set of actions given the model structure. Assume that a sequence of M observations for a given user i is $\boldsymbol{Y}_i = Y_{i1}, Y_{i2}, \ldots, Y_{iM}$, where $Y_{im} \in \mathcal{Y}, m \in \{1,2,\ldots,M\}$. Also, assume that $\boldsymbol{Y}_i$ is the result of a sequence of latent states, $\boldsymbol{S}_i = S_{i1}, S_{i2}, \ldots, S_{iM}$, where $S_{im} \in \mathcal{S}$, with respective input vectors $\boldsymbol{X}_{i1:M-1} = \boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{iM-1}$. Figure 2 illustrates these sequences along with their interactions.

Based on the model structure, the conditional likelihood of observing $\boldsymbol{Y}_i$ is

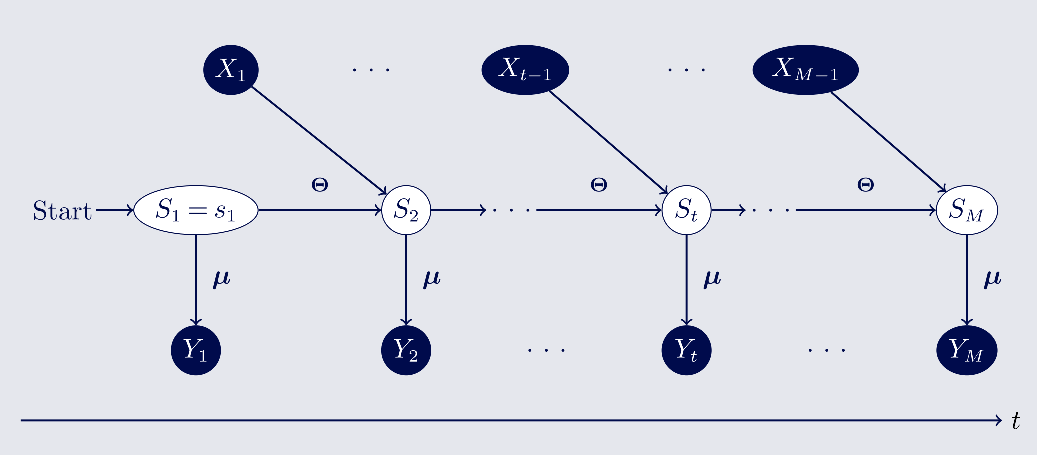
$$\Pr\left(\boldsymbol{Y}_i | \boldsymbol{S}_i; \boldsymbol{\mu}\right) = \prod_{t=1}^{M} \Pr\left(Y_{it} | S_{it}; \boldsymbol{\mu}\right) = \prod_{t=1}^{M} \mu^{S_{it}}_{Y_{it}}. \tag{7}$$

The conditional probability of getting the sequence $\boldsymbol{S}_i$ is

$$\begin{aligned} &\Pr(\boldsymbol{S}_i | \boldsymbol{\Theta}, \boldsymbol{X}_{i1:M-1}) \\ &\quad = \pi(S_1) \prod_{t=2}^{M} \Pr\left(S_{it} | S_{it-1}; \boldsymbol{\Theta}, \boldsymbol{X}_{i1:M-1}\right) = \prod_{t=2}^{M} \lambda^{S_{it-1},S_t}_{\boldsymbol{\Theta}\boldsymbol{X}_{it-1}}, \end{aligned} \tag{8}$$

where $\pi(S_1) = \pi(S_1 = s_1) = 1$ because all users deterministically begin in state s_1.

Figure 2. (Color online) The HMM-AFT Framework over Time



Notes. Survival models define transition probabilities in the latent-state model that describes user engagement. The covariate vector $\boldsymbol{X}_{t-1}$ and the parameter vectors $\boldsymbol{\Theta}$ affect the transitional probabilities from state S_{t-1} to S_t, $S_{t-1}, S_t \in \mathcal{S}$. Parameter vector $\boldsymbol{\mu}$ defines the probabilities for observing $Y_t \in \mathcal{Y}$. Latent states are in clear ellipses, whereas observed actions are shaded.

Based on this analysis and Figure 2, the likelihood of this sequence of observations for user i is as follows:

$$\begin{aligned} l(\boldsymbol{Y}_i;\boldsymbol{\Theta},\boldsymbol{\mu},\boldsymbol{X}_{i1:M-1}) &= \Pr(\boldsymbol{Y}_i|\boldsymbol{\Theta},\boldsymbol{\mu},\boldsymbol{X}_{i1:M-1}) \\ &= \sum_{\forall \boldsymbol{S}_i} \Pr(\boldsymbol{Y}_i,\boldsymbol{S}_i|\boldsymbol{\Theta},\boldsymbol{\mu},\boldsymbol{X}_{i1:M-1}) \\ &\overset{\text{Figure 2}}{=} \sum_{\forall \boldsymbol{S}_i} \Pr(\boldsymbol{Y}_i|\boldsymbol{S}_i;\boldsymbol{\mu}) \Pr(\boldsymbol{S}_i|\boldsymbol{\Theta},\boldsymbol{X}_{i1:M-1}) \\ &\overset{\text{Equations 7,8}}{=} \mu^{S_{i1}}_{Y_{i1}} \sum_{\forall \boldsymbol{S}_i} \prod_{t=2}^{M} \mu^{S_{it}}_{Y_{it}} \lambda^{S_{it-1},S_t}_{\boldsymbol{\Theta}\boldsymbol{X}_{it-1}}, \end{aligned} \qquad (9)$$

where the structure of the HMM allows decomposition of the joint probability of $\Pr(\boldsymbol{Y}_i, \boldsymbol{S}_i|\boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{X}_{i1:M-1})$ (Murphy 2012). Finally, the complete likelihood for N users is

$$L(\boldsymbol{\Theta},\boldsymbol{\mu}) = \prod_{i=1}^{N} l(\boldsymbol{Y}_i;\boldsymbol{\Theta},\boldsymbol{\mu},\boldsymbol{X}_{i1:M-1}). \qquad (10)$$

For efficient estimation of the parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\mu}$ that maximize this complete likelihood, we use the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm (Byrd et al. 1995). (Alternatively, minimizing an error function can estimate parameters (Kokkodis 2019a).)

3.2. Predicting Community Welfare Through User Engagement

A community's welfare depends on highly engaged users (Section 2.5). Hence, given that the HMM-AFT framework predicts the state of engagement for each user at any point in time, it can also provide information that predicts a community's future welfare.

We demonstrate this by designing a predictive model for the multiple dimensions of community welfare. Section 2.5 highlights five measures that reflect a community's welfare: (a) the total number of new threads (W_1), (b) the total number of responses (W_2), (c) the ratio of the number of responses per thread (W_3), (d) the mean number of unique users per thread (W_4), and (e) the total number of reciprocal posts (W_5). The first two measures capture user activity (Ridings et al. 2006, Saltz et al. 2007, Shen and Khalifa 2007, Himelboim et al. 2009, Millington 2012, Schneider et al. 2013), the next two capture the interactivity between users (Preece 2001, Viégas and Smith 2004, Angeletou et al. 2011, Bernstein et al. 2011, Cheng et al. 2015), and the last measure captures reciprocity (Goldstein et al. 2001, Molm et al. 2007, Haines et al. 2011). We extend these by adding three new measures that capture the ability of the community to attract and engage new users: (f) the number of new users (W_6), (g) the number of new contributors (W_7), and (h) the percentage of new users who contribute (W_8). Together, the eight measures constitute a holistic view of the welfare of a community.

We evaluate five different predictive models: ϵ-Support Vector regression model (ϵ-SV) (Shevade et al. 2000, Sapankevych and Sankar 2009), an autoregressive integrated moving average model with explanatory variables (ARIMAX) (Friedman and Meiselman 1963), a k-nearest neighbor regression (kNN), recurrent neural networks (long short-term memory, or LSTM) (Hochreiter and Schmidhuber 1997), and gradient boosting (XGBoost) (Chen and Guestrin 2016). All models use the same vector of predictors $\boldsymbol{Z}$, which includes (1) variables that account seasonal effects, (2) time-lagged versions of the five welfare measures $W_1 - W_5$, (3) time-lagged variables that encode the community's evolving size, and (4) variables based on the latent states of user engagement learned via the HMM-AFT framework (HMM variables $\mathscr{A}$).

4. Research Context

Our empirical analysis focuses on a large online community for diabetes patients, DiabetesForum, that began in June 2007. The site maintains forums where users can discuss issues with other users. Additionally, it offers a variety of firm-created resources (such as nutrition guides, product information, video interviews, etc.), sponsors offline events, and hosts webinars. The community directs the initiatives toward creating a positive and supportive environment to promote health for those living with diabetes.

4.1. Overview

The data set from the DiabetesForum community includes the activity of all 45,308 users who joined between June 2007 and October 2017. Users can create new threads to ask questions or make observations that they believe will interest the community. Other users can then respond to the new threads (or prior responses) as well as indicate support for existing content (by clicking on a "Like" indicator). Thread content and the number of "likes" are visible to all users. Within the data set, users generated a total of 49,904 new threads and 483,810 responses to these new threads. Figure 3 illustrates community activity over time. The community reaches a peak of responses and new threads just before 200 weeks. New content generation becomes sparser after that time. Furthermore, the number of new threads and the number of responses are correlated.

DiabetesForum is an insightful empirical context for research on online communities for several reasons. First, the community maintained a consistent structure, thus reducing the potential influence of exogenous shocks due to changes in the technology platform. Second, its tenure allows sufficient time (10 full years) to observe multiple phases of growth and decline and therefore, exhibit variance in user engagement. Third, the community focus on diabetes attracts users with an intrinsic interest in topics that are likely to be deeply important to them. Fourth, DiabetesForum fits the general contribution pattern of other online communities (i.e., the 1-9-90 principle (Van Mierlo 2014)), with only a small percentage of users contributing content (Figure 4). Even further, this percentage declines as the total number of registered users increases over time.

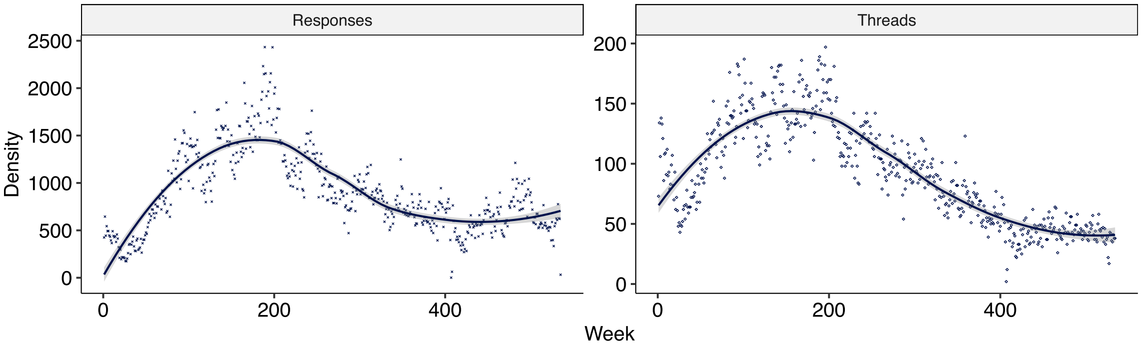
4.2. Variables That Affect Transitions

The rich DiabetesForum data provide several measures of community aspects that may be associated with user engagement (e.g., user-to-user interactions, observable user actions, and the users' topical interests). The covariate vector $\boldsymbol{X}_t$ includes these measures.

4.2.1. Community Interactions. Community feedback is important to user confidence and engagement (Lampe and Johnston 2005, Joyce and Kraut 2006, Moon and Sproull 2008, Ozturk and Nickerson 2015). In the DiabetesForum context, feedback is directly encoded in the endorsements (likes) that the users receive for their contributions by other community members as well as through community badges. We thus measure direct feedback through three variables: the average number of likes that each user receives per post ("Likes received (average per post)"), the total number of unique users who liked the focal user's contributions ("Received likes (unique users)"), and whether the user has received a badge ("Badges").

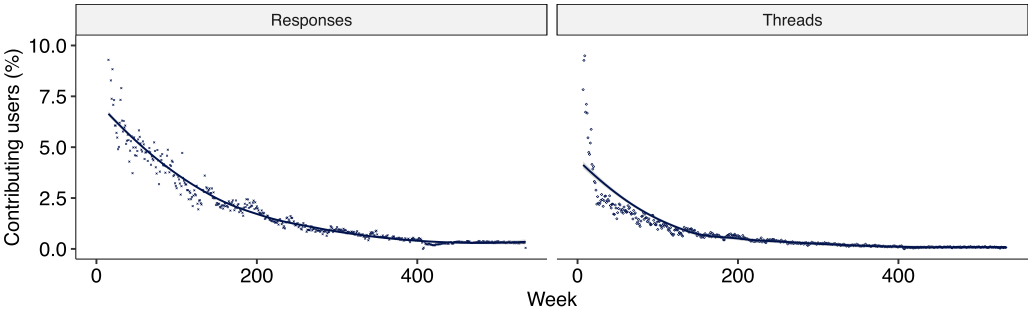
Community interactions can also provide indirect feedback. For example, a user who creates a well-received thread or response that is followed by a lengthy discussion with other users can be encouraged to get more engaged with the community. To measure this type of indirect feedback, we compute the number of responses in a thread after a user's

Figure 3. (Color online) Overview of Community Activity



Note. User activity (new threads, responses to existing threads) from June 2007 until October 2017.

Figure 4. (Color online) Contributing Users



Note. Like many other online communities, most users are not actively engaged.

response (“Responses after response”) and the average number of responses that a user’s thread receives (“Avg responses per thread (received)”). Both of these types of feedback could affect the user’s confidence.

4.2.2. Observable User Actions. Additionally, we measure activities by the user that are directly observable. For example, the number of times that a user responds to any thread (“Number of responses”), the number of unique threads a user responds to (“Unique topics responded”), the number of new threads that a user creates (“Number of threads”), and the number of unique users who a user likes (“Gave likes (unique users)”) represent the basic participatory actions for community members (Huffaker 2010). Remaining active for an extended period of time correlates positively with the contribution quality, whereas being intermittently active presents a weak but negative correlation (Nam et al. 2009). We capture such patterns via the average number of weeks between consecutive user actions (i.e., a thread creation or a new response), the standard deviation of this quantity (“Weeks between actions,” average and standard deviation), and the number of weeks that have passed from a user’s last action (“Weeks from last action”). Additionally, early engagement for community users is important (Arguello et al. 2006, Burke et al. 2009). We directly observe the initial actions of each user, such as whether the user has created a thread in the first week after joining the platform (“Thread in first week”) and whether the user has responded to a thread in the first week after joining the platform (“Response in first week”). Furthermore, we observe the state of each thread at the time when a user chooses to contribute to it. For instance, a user might choose to contribute in a fairly new thread, with few responses. To the contrary, a different user might choose to contribute to a fairly mature thread, with many responses. Each of those actions might have an effect on the subsequent user level of engagement with the community because different thread maturity levels might result in different community interactions and as a result, in varying feedback to the user (Moon and Sproull 2008). We measure the maturity level of a thread through the number of responses before a user’s choice to respond (“Responses before response”). Finally, we indicate whether a user has a profile picture (“User image (binary)”), which is correlated with higher levels of engagement (Adaji and Vassileva 2016).

4.2.3. Topical Interests. Users who have broader interests disseminate information within the community (Hecking et al. 2015). The average number of responses per thread (“User avg responses per thread”) measures the broadness of user interests. By measuring broadness, we control for potential “superposters” (Huang et al. 2014). Furthermore, beyond just the number of responses per thread, we also mine the content that each user contributes. User engagement may vary by their topical interests (i.e., more/less engaged users favoring certain topics). We estimate topical interests through deep learning (“Deep learning features”). Specifically, we create user-specific documents by concatenating the posts that each user contributes. We then use the distributed memory model (Le and Mikolov 2014) to embed each user into a multidimensional semantic space according to their respective document. Online Appendix A describes the detail behind this method.

Table 2 summarizes the variables extracted from the DiabetesForum community (excluding the deep learning features). Online Appendix B contains the correlogram for these variables. We log-transform variables with long tails and standardize all variables for faster convergence.

4.3. Welfare Variables

To evaluate the performance of the three models that predict welfare (Section 3.2; ϵ-SV, ARIMAX, and kNN), we define a vector of predictive variables $\boldsymbol{Z}$. The focal predictors are the variables based on the users’ latent states of engagement as learned by the

Table 2. Descriptive Statistics of Community Activity

	Mean	Median	Min	Max	Standard deviation
Transition variables					
User actions					
Number of threads (count)	1.10	0	0	970	8.56
Number of responses (count)	10.68	0	0	9,021	126.53
Response in first week (binary)	0.16	0	0	1	0.37
Thread in first week (binary)	0.09	0	0	1	0.29
Weeks between actions (count)	13.41	1	1	433	32.73
Gave likes (unique users)	0.06	0	0	68	0.93
User responses (average per thread)	0.30	0	0	31	0.95
Unique topics responded (count)	2.03	0	0	3,777	47.31
Responses before response (count)	4.70	0	0	1,482	20.68
User image (binary)	0.35	0	0	1	0.47
Weeks from last action (count)	6.25	1	0	499	20.99
Community actions					
Likes received (average per post)	0.01	0	0	10	0.18
Received likes (unique users)	0.03	0	0	20	0.38
Responses after response (count)	2.36	0	0	987	15.07
Responses received (average per thread)	0.49	0	0	161	2.62
Badges (binary)	0.05	0	0	1	0.21
Welfare variables					
W_1: threads (count)	281.51	277	7	532	124.76
W_2: responses (count)	2,713.48	2,484	40	6,803	1,270.45
W_3: responses per thread (count)	5.56	5.61	1.78	9.99	1.29
W_4: unique users per thread (count)	3.71	3.74	1.66	5.31	0.56
W_5: reciprocity (count)	473.95	443	11	1,051	187.60
W_6: new users (count)	84.93	88	4	183	34.59
W_7: new contributors (count)	17.95	17	1	39	9.40
W_8: new users who contribute (%)	0.13	0.11	0	0.45	0.07

Note. The data include 483,810 responses in 49,904 threads by 45,308 users from June 2007 until October 2017.

HMM-AFT. Specifically, for each period p, we compute the following set $\mathcal{A}$ of HMM variables:

- the number of users who began p as lurkers and became engaged during p,
- the number of users who began p as engaged and remained engaged throughout p, and
- the number of users who began p as engaged and regressed to lurkers during p.

These variables capture the transitions (from lurker to highly engaged and from highly engaged to lurker) that occur during a period p as well as the number of users who remain highly engaged throughout p. In addition, time-lagged versions of these variables incorporate the consistency of such events during the community's recent history. Intuitively, a community that enjoys consistently high numbers of transitions to more engaged states and low numbers of transitions to less engaged states for extended periods of time is more likely to improve its future welfare.

To control for the increasing number of users as the community grows, we create the following two variables for each period p:

- the number of users who registered before p and
- the number of users who registered during period p.

Because we focus on timeseries prediction of welfare measures, we further include time-lagged variables of the eight welfare measures $W_1 - W_8$ (Section 3.2). Finally, we control for trending and seasonal effects through a time trend variable for each year as well as dummy variables for the month and quarter of the year. The bottom of Table 2 shows the descriptive statistics of the eight welfare measures. Table 3 describes all of the welfare variables.

5. Results

Section 5.1 describes the results of estimating latent states of user engagement. This process includes choosing an underlying survival function f as well as the total number of states K. Section 5.2 models the community using the HMM-AFT framework. Then, Section 5.3 discusses the performance of the HMM-AFT framework in modeling and predicting user contribution and community welfare. Finally, Section 5.4 shows how the HMM-AFT framework can inform managerial intervention. For the HMM-AFT, we aggregate user activity at the weekly level. Experimentation with daily and monthly aggregations resulted in qualitatively similar frameworks. Online Appendix D shows that the HMM-AFT approach generalizes to five additional online communities,

Table 3. Predictors of Community Welfare (Vector $\boldsymbol{Z}$)

	Variable
Observed community variables	Community welfare measures ($W_1 - W_8$)[a]
Observed community variables	Time trend (annual)
Observed community variables	Quarter dummies
Observed community variables	Monthly dummies
Observed community variables	Users who registered before p (count)[a]
Observed community variables	Users who registered during p (count)[a]
HMM variables ($\mathscr{A}$)	Users who began p as lurkers and became highly engaged during p (count)[a]
HMM variables ($\mathscr{A}$)	Users who began p as engaged and remained engaged throughout p (count)[a]
HMM variables ($\mathscr{A}$)	Users who began p as engaged and regressed to lurkers during p (count)[a]

[a] The vector of predictors $\boldsymbol{Z}$ includes time-lagged versions for each of these variables.

including DronesForum (for drone enthusiasts), CookingForum (for food lovers), DietForum (for males interested in healthy lifestyles), AlternativeDietForum (for meal replacement enthusiasts), and CompeteForum (to encourage technological development).

5.1. Participation Types and Parameter Estimation

The HMM considers five types of user contribution ($\mathscr{Y}$): "Lurk," "Append," "Respond," "Ask," and "Share." Most of these actions are directly observable. To identify whether a new thread is an "Ask" or a "Share" post, we run latent dirichlet allocation (Blei et al. 2003). Online Appendix C provides the details of this process and examples of "Ask" and "Share" posts. When a user takes multiple actions within a week, we map those to the one associated with the higher level of self-confidence. For instance, if a user both responds and asks a new question during the same week, we observe "Ask." (Modeling more actions creates sparse data sets and lower predictive performance.)

The HMM-AFT framework requires two choices: the number of states (K) and the survival function (f). To select, we compare configurations and calculate their Bayesian information criterion (BIC) scores (Schwarz 1978, Murphy 2012). In particular, the framework considers the following continuous probability distributions for function f:

$$f \in \{\text{Exponential, Loglogistic, Lognormal, Weibull}\}, \tag{11}$$

and the following set of number of states:

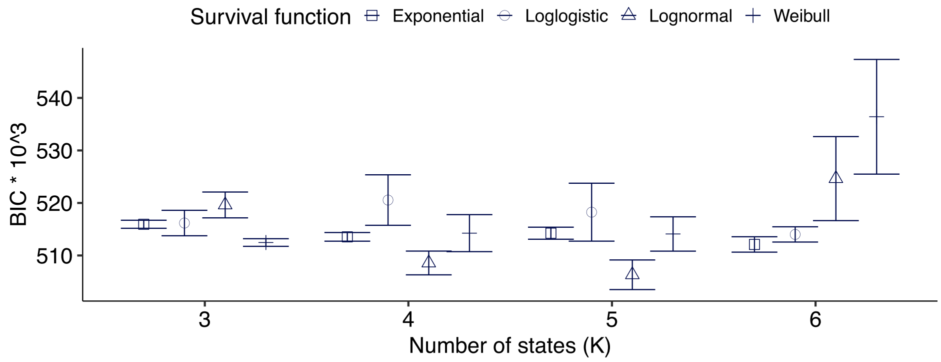
$$K \in \{3, \ldots, 6\}. \tag{12}$$

For each combination in $\{K \times f\}$, we estimate the parameters $\boldsymbol{\Theta}, \boldsymbol{\mu}$ that maximize the likelihood of Equation (10). Maximizations such as these are prone to finding local maxima rather than the global maximum because of the initial parameters. Hence, to increase the likelihood of selecting the optimal number of states K, we search 1,000 randomly generated initial parameters for each combination in $\{K \times f\}$. Figure 5 shows the BIC score (Schwarz 1978) for these configurations. The log-normal function with five states ($K = 5$) yields the lowest BIC score.

5.2. State Transitions

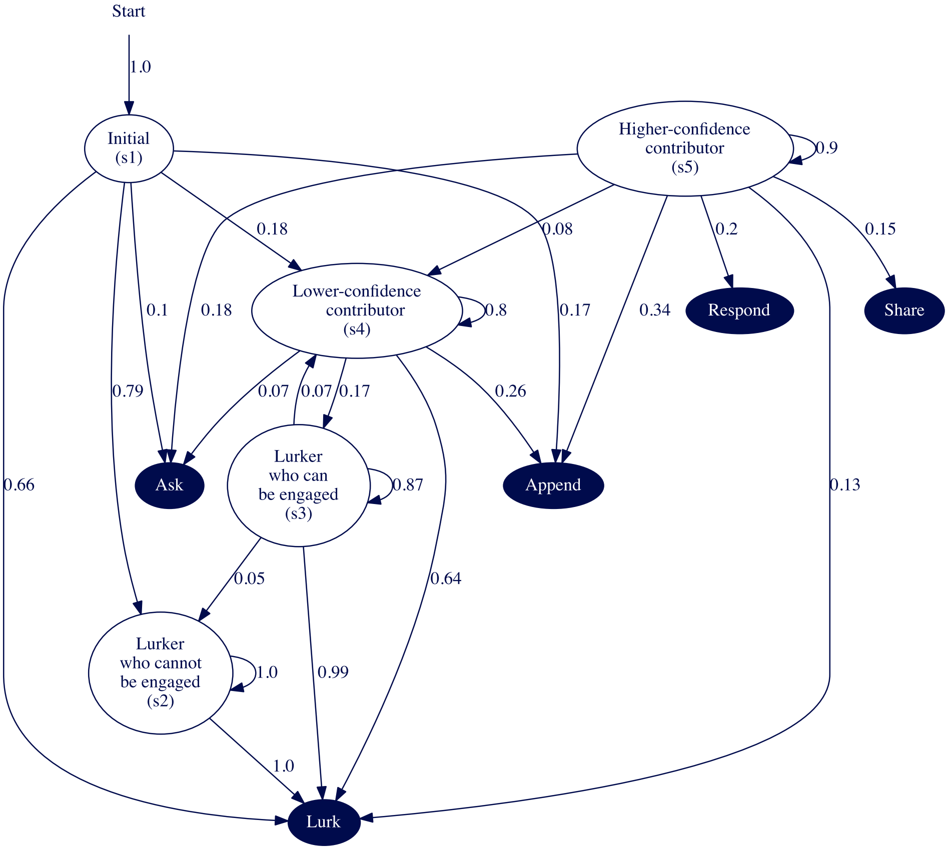
Using these optimal parameters, Figure 6 shows the resulting HMM-AFT framework with latent (transparent) states and observable (filled) actions (Koller and Friedman 2009, Murphy 2012). Each latent state has a different probability distribution across all

Figure 5. State and Distribution Function Selection



Note. Error bars show 95% confidence intervals.

Figure 6. (Color online) Latent-State Transitions Within the DiabetesForum Community



Notes. The HMM structure that yields the lowest BIC score for $K = 5$. Similar to Figure 1, s_1 is in the top layer (starting state), and s_2 to s_5 are in the bottom layer. (For increased readability, we only show transitions and emissions with probability greater than 0.05.)

possible actions from $\mathcal{Y}$ (i.e., $\mu_y^{s_k}$, $y \in \mathcal{Y}$, and $s_k \in \mathcal{S}$). For example, a user in the initial state s_1 has on average 0.66 probability to "Lurk," 0.17 probability to append to current threads, 0.01 probability to ask a question, 0.04 to share information, and only 0.02 chance to be the first responder to a thread.

Figure 6 also shows the distinction of the two HMM layers: the first layer contains the initial state s_1, where all users deterministically begin when they first join the platform. From s_1, users stochastically transition to the second layer (i.e., states s_2 to s_5). More importantly, the graph shows a clear separation between different types of lurkers—lurkers who cannot be engaged (state s_2) versus lurkers who can become active again (state s_3). States s_4 and s_5 are the higher engagement states, which we distinguish into "Lower confidence" and "Higher confidence." In the "Lower-confidence" state, users mainly append responses to current threads (26% chance). In the "Higher-confidence" state s_5, users ask new questions (18%), share information (15%), and are the first to respond (20%). Overall, state s_5 represents the most active contributors of the community.

The emission probabilities might appear low at first. However, because the HMM is trained over sequences of weekly observations, the cumulative effect of these probabilities increases considerably. For instance, for a user who is in state s_5, the probability of not contributing at all after 4 weeks would be practically zero (i.e., $0.13^4 = 0.0003$). Beyond just the likelihood of contributing, the model also indicates that a user who is in the same state s_5 for 4 weeks will ask on average $4 \times 0.18 \sim 0.72$ new questions, share $4 \times 0.15 = 0.6$ new information threads, and respond $4 \times 0.34 = 1.36$ times (expected mean of a sequence of Bernoulli trials; i.e., a binomial distribution).

This structure further allows observation of hidden patterns of user evolution. One week after joining the platform, most users end up in the predominantly unengaged state s_2. Around 18% transition to the "Lower-confidence" state s_4, where they start contributing. Only 3% of the users end up in the "Higher-confidence" state after their first week of joining the platform. From state s_5, users will likely keep contributing, but around 8% will transition to the "Lower-confidence" state; from there, few (17%) will end up in state s_3. After they are in s_3, users will likely stay there, with some of them (7%) reactivating and moving to state s_4.

5.3. Predicting Contribution and Community Welfare

The HMM-AFT framework allows managers to predict individual user engagement (Section 5.3.1) and future community welfare (Section 5.3.2). Next, we compare its predictive performance against previous works and other baselines.

5.3.1. Predicting Contribution. One of the purposes of the HMM-AFT framework is to provide tools that help community managers effectively forecast current and future user engagement. A complete framework needs to accurately predict all types of contribution (i.e., "Respond," "Ask," "Share"). Such accurate predictions allow managers to make type-specific interventions. For instance, managers can target users who are more likely to "Respond" to new or unanswered questions, users who are more likely to "Append" to open discussion threads, and so forth.

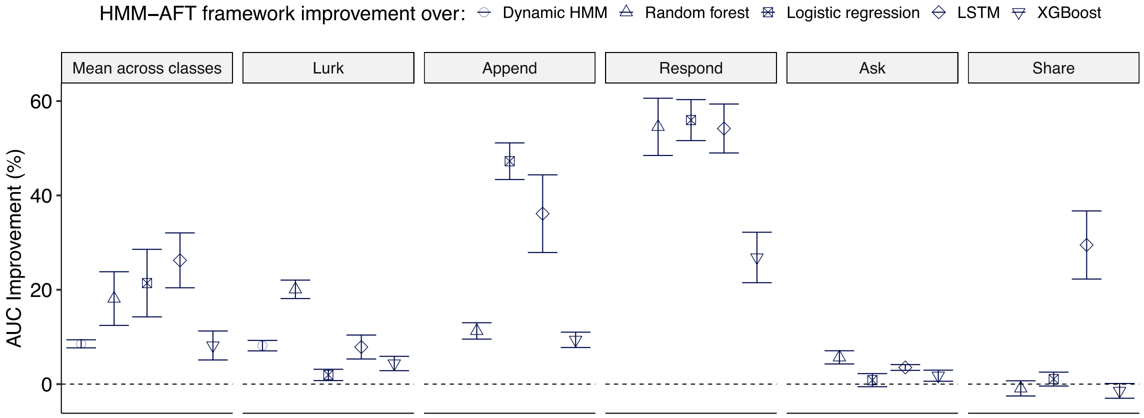
To predict individual user contribution, we allow some of the variables of vector $\boldsymbol{X}$ to affect emissions. Furthermore, because this adaptation increases significantly the number of parameters that we need to estimate, we use step forward feature selection (Ferri et al. 1994). To benchmark the performance of the HMM-AFT, we compare with several advanced alternative algorithms. (Online Appendix E shows the details of these implementations along with the respective parameter tuning.) These include the following models.

- Static models. Several static models estimate the probability of each type of contribution (multiclass classification). Specifically, we consider logistic regression, random forest, and gradient boosting for classification (XGBoost 2018).
- Dynamic models. Because users are dynamic entities with correlated observations over time, we consider two dynamic approaches: LSTM network (Hochreiter and Schmidhuber 1997) and an HMM (dynamic HMM) that estimates the probability of contributing a single type of contribution (one or more responses) (Chen et al. 2018).

Figure 7 shows the 10-fold crossvalidated average and per class area under the curve (AUC) improvement scores of each approach. (AUC scores across all classes and models range between 51% and 92%, with an average of 77%.) The HMM-AFT approach significantly ($p < 0.001$) outperforms all other approaches by an *average* AUC improvement between 9% and 26% ("Mean across classes" plot in Figure 7). For per class accuracy, the HMM-AFT approach significantly ($p < 0.001$) outperforms all alternative predictive models for classes "Lurk," "Append," and "Respond." For predicting "Ask," the HMM-AFT approach outperforms logistic regression ($p < 0.1$) and all other approaches ($p < 0.001$). For predicting "Share," the HMM-AFT approach performs on par with logistic regression, random forest, and XGBoost models and significantly ($p < 0.001$) outperforms the LSTM model. Overall, the results reveal that, by learning a probabilistic state space that accurately encodes user behavior, the HMM-AFT can generalize and deliver accurate predictions of user participation as measured on unseen testing data. (Figure A.11 in the online appendix shows how the HMM-AFT approach outperforms all other approaches in five additional communities.)

5.3.2. Predicting Community Welfare. Next, we examine how knowledge of the users' latent states relate to overall community welfare (Section 3.2). Five predictive models (ϵ-SV, ARIMAX, kNN, LSTM, and XGBoost) estimate the value of the welfare dimensions W_l, for $l \in \{1, \ldots, 8\}$. We use the set of predictor and lagged variables (Section 4.3), aggregating the values of the predictors by month. The focus on monthly aggregates allows us to capture more information within each time unit. (Analysis with weekly data led to worse out-of-sample performance because of

Figure 7. Forecasting User Contribution



Notes. The 10-fold crossvalidated AUC scores for each approach. Error bars represent 95% confidence intervals.

sensitivity to outliers. Monthly aggregates smooth out outlier effects by considering a longer part of the timeline.)

Three baselines benchmark the performance of the HMM-AFT.

- No state variables does not consider the HMM variables ($\mathscr{A}$).
- Dynamic HMM includes user state information (Chen et al. 2018).
- Dynamic network uses lurker ranking information (Tagarelli and Interdonato 2014). For the required social network graph, we assume edges between users who interact in the same thread.

All models consider time-lagged versions of every variable except for trend and seasonal effects (Section 4.3). Following best practices for time series forecasting, we select the number of lags by using out-of-sample prediction with the root mean squared error (RMSE) as an objective function (Hyndman and Khandakar 2007). After experimenting with multiple alternatives ($\{6, 12, 18, 24\}$ lags), we found that 12 lags led to the best results for all three models.

For all models, we tune their respective parameters using a grid search on a holdout sample. We train the models using the first 36 months of data and test their efficacy on the rest. In the testing step, the models simultaneously predict each welfare measure up to 12 months in the future. Predictions occur sequentially. To predict r months ahead (i.e., prediction step = r), we include a sliding window of the 12 prior months as lagged variables. Formally, let $q, \widehat{q}^{\,r}$ be vectors that include the actual values (q) and the corresponding "r steps ahead" predictions ($\widehat{q}^{\,r}$) of a welfare measure for step r. Then, the root mean squared error at step r ($RMSE_r$) is

$$RMSE_r(q, \widehat{q}^{\,r}) = \sqrt{\frac{\sum_{j=1}^{J}\left(q_j - \widehat{q}_j^{\,r}\right)^2}{J}},$$

where J is the total number of predictions in the test data for r steps ahead.

Figure 8 shows the average improvement of all models over the "No state variables" approach. Almost all models perform better than the "No state variables" approach for all welfare measures. Exceptions include the dynamic network approach (Tagarelli and Interdonato 2014) for W_3, W_4, W_8 and the dynamic HMM approach (Chen et al. 2018) for W_7. Across all models (ARIMAX, kNN, and ϵ-SV), the HMM-AFT framework significantly outperforms all other approaches.

5.4. Simulating Managerial Intervention

Simulations assess how well managerial interventions could benefit from the predictive performance of the HMM-AFT. We begin by creating the monthly time series for each of the eight welfare measures. Using the series for each measure, the ϵ-SV algorithm predicts the welfare values for the following 12 months. Then, for a given welfare measure l, we define the total predicted change as

$$\text{change} - \text{sum}_l = \sum_{i}^{1} \hat{W}_l^{i+1} - \hat{W}_l^{i}, \tag{13}$$

where $\hat{W}_l^i$ is the ith prediction of welfare measure W_l. Conceptually, a lower change sum represents a period with a greater reduction in welfare. We then rank all of the predicted 12-month periods by their respective change sum scores in ascending order. Thus, the top-ranked period is the one with the lowest change sum. After repeating this for all eight welfare measures, we identify the 12-month interval (in the test set) with the highest mean rank across the eight measures. Conceptually, this period is the time during which the community suffered the lowest change sum as aggregated across welfare measures.

We repeat this process for each of the benchmark approaches and the HMM-AFT. For each approach, we compute its mean rank according to the actual (not predicted) series of the eight welfare measures. Table 4 shows the results. The three benchmark approaches fail to identify the most severe drop in welfare measures, with the dynamic HMM approach doing better than the other two, identifying the 18th worst actual period. To the contrary, the HMM-AFT identifies the second worst interval between weeks 191 and 232. Figure 3 shows the drop in contribution during this period, demonstrating how greater predictive performance can guide managerial interventions.

To simulate a realistic intervention during the predicted interval (i.e., starting on week 190), we first assume that the community managers have a limited budget on how many users they can target. We consider budgets of 2.5% and 5% of the total community users. The HMM-AFT indicates a target group that the community should focus on lurkers who can be engaged (state s_3). We consider two targeting strategies:

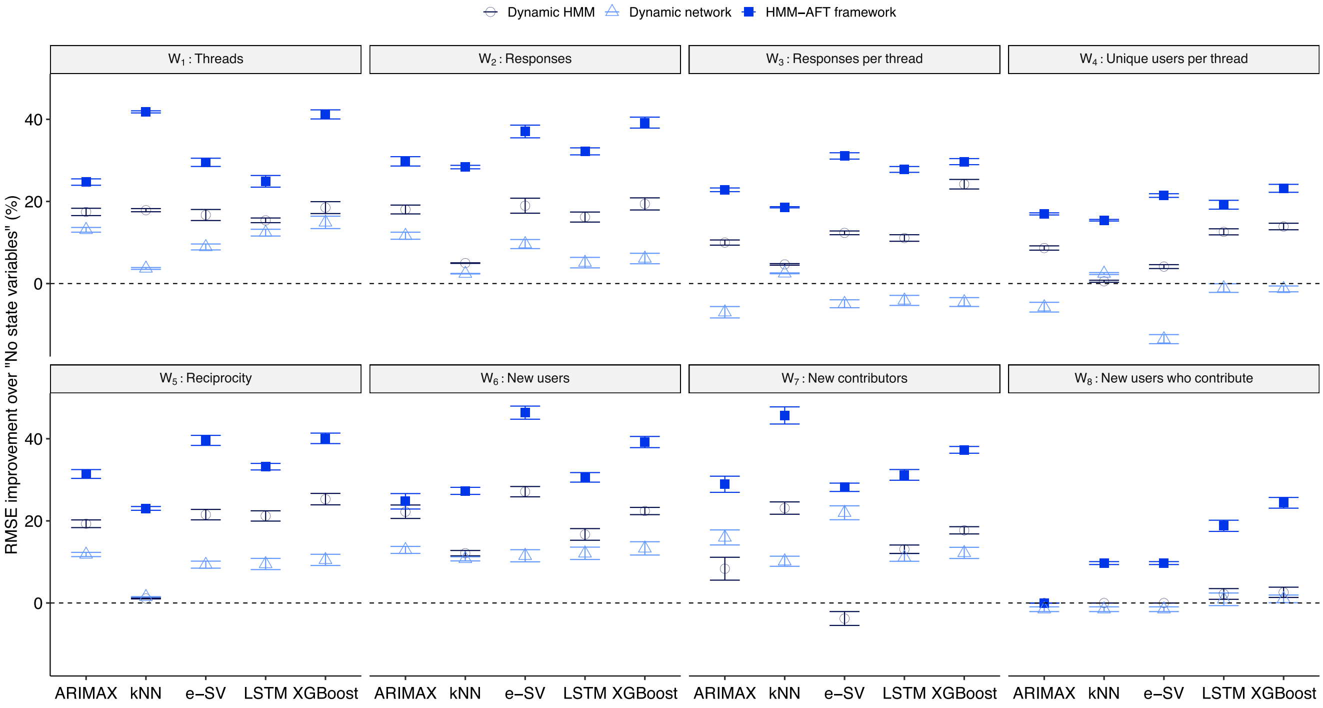
- strategy 1, random budget allocation to all users; and
- strategy 2, targeted budget allocation to lurkers who can be engaged (state s_3).

We further assume an intervention efficacy $IE \in \{5\%, 10\%, 15\%\}$. Hence, a user i in state s_k has probability to transition to a high activity state ($s_l \in \{s_4, s_5\}$) according to the following:

$$\Pr(S_{t+1} = s_l | S_t = s_k) = T(\mathbf{\Theta}, \boldsymbol{X}_t)[k][l] + IE\ , \forall k \neq l. \tag{14}$$

If treated users transition to a high activity state, we simulate their contributions over the next weeks according to the HMM-AFT framework. We repeat

Figure 8. (Color online) Welfare Prediction



Notes. The predictive performance of the dynamic HMM (Chen et al. 2018), the dynamic network (Tagarelli and Interdonato 2014), and the HMM-AFT framework compared with a no state variables approach. In all welfare metrics, the HMM-AFT significantly outperforms all three baselines. Error bars represent 95% confidence intervals.

this process 100 times for each combination of budget, intervention efficiency, and targeting strategy.

Figure 9 shows the average increased contribution under the two intervention strategies in terms of both responses and threads. The increased contribution is averaged over the interval of intervention (weeks 191–232.) Targeting lurkers who can be engaged (state s_3) outperforms the alternative strategy of targeting users randomly. Interventions could result in up to a 12-month average of 18% increase in user contribution depending on how efficient they are.

These results show the utility of monitoring and predicting user engagement events in an online community. Engagement states not only provide a better theoretical understanding of online communities, but they also have natural applications for community managers and their efforts to enhance their community's welfare.

Table 4. Predictions of the Most Severe Welfare Change

Approach	Interval (weeks)	Interval (months)	Actual rank
No state variables	323–365	80–91	54
Dynamic HMM	255–297	63–74	18
Dynamic network	331–373	82–93	58
HMM-AFT	191–232	47–58	2

Note. The benchmark models fail to predict the critical period between weeks 191 and 232.

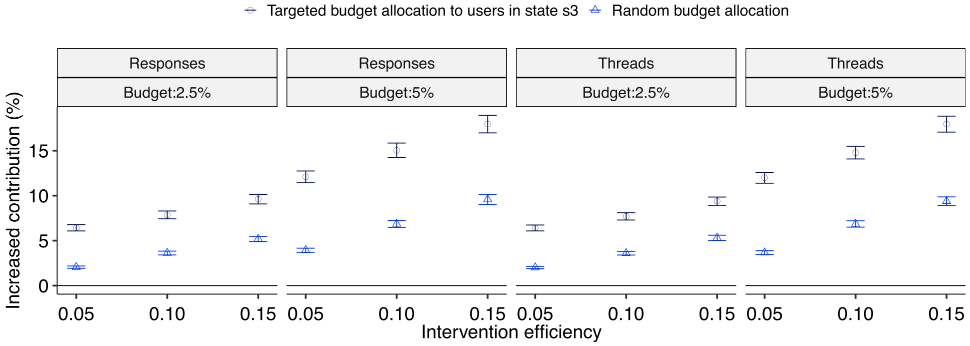
6. Discussion

In online communities, only a small fraction of the users typically contribute content, whereas the vast majority passively lurk. Our formal framework (HMM-AFT) allows us to study the transition of these less active users (lurkers) to greater engagement by modeling user evolution through a latent space of engagement derived from observed participatory actions. The HMM-AFT framework builds off of the extensive theoretical work on user participation and engagement, which motivates us to construct a model that differentiates between observable (and possibly incidental) actions from latent states that better reflect a user's level of engagement. Although we focus on a single community (DiabetesForum), we illustrate the generalizability of the framework and theoretical predictions using five other diverse platforms.

6.1. Contributions to Research

Although many previous efforts to model online community activity focus on observable user actions, latent approaches offer an opportunity to better understand these communities. For example, observable actions may be incidental or reflect ephemeral bursts of activity and may not reflect a user's engagement with community. To address potential limitations in the use of observed actions, researchers often use thresholds (Brzozowski et al. 2009, Healey et al. 2014,

Figure 9. (Color online) Simulating Intervention



Notes. Percentage of increased contribution is averaged over the 12-month interval of intervention. Depending on the assumption regarding the efficiency of the intervention (*IE*), there is an increase in contribution. Targeting lurkers who can be engaged (state s_3) results in higher contributions compared with targeting users across all available states.

Olteanu et al. 2016). Recent work considers engagement as a latent state in formal frameworks (Chen et al. 2018). We extend these approaches by deeper consideration of those passive users who have limited evidence of engagement with the community, the largest segment of the user population. Our formal framework models both observed participatory actions and latent engagement states for *all* types of users without any thresholds or assumptions on their current level of engagement. The HMM-AFT framework enables us to reveal unique community-specific patterns of user evolution in a diverse set of online platforms (Figure 6 and Figures A.14–A.17 in the online appendix). Future research can extend this framework to incorporate additional information and further improve the modeling of online communities.

Through dynamically modeling engagement according to user confidence, this work is the first to clearly differentiate lurkers who can be engaged in the future from those who cannot. This distinction is only feasible through the appropriated coding of types of contribution. Future research can use our framework to study user engagement through confidence-driven participation.

Finally, we show that the latent states of engagement are useful in predicting future community welfare. Using multiple measures of community welfare, the latent approaches not only increase the understanding of individual user engagement but also, improve the ability to understand future community welfare. Improved prediction accuracy of multiple models that include latent-state variables illustrates the additional benefit of the latent approach. Communities with an increased number of engaged users are more likely to achieve higher welfare values in the future. Future research can use this approach to better design and intervene at the individual, community, and platform levels relative to future community welfare.

6.2. Contributions to Practice

The proposed design guides practitioners to address challenges in modeling user participation and engagement in an online community.

- Modeling latent states. A two-layer HMM structure provides a more accurate representation of patterns of participation in an online community than a standard HMM (Section 3.1).
- HMM architecture. A series of observed characteristics shapes transitions between latent states of varying user engagement (Section 3.1).
- Parameter estimation. Parameter estimation includes the derivation of the global likelihood of the model and the estimation process of all of the parameters (Sections 3.1 and 5.1).
- Distribution choices and states selection. Distribution choices and states selection include the choice of appropriate underlying distributions and number of states that best fit the particular context (Section 5.1).
- Designing variables and identifying contribution types. New variables designed specifically for the online community context (e.g., the number of responses before and after an action of a user and the deep learning variables) are informative predictors of user engagement and can improve the performance of future modeling and prediction tasks in this context (Section 4.2). Furthermore, categorization of contribution types through topic modeling allows an efficient separation of users according to their confidence level.

6.3. Managerial Implications

Managerial action (including different types of user rewards, recognition badges, increased privileges, etc.) can increase user engagement with the community (Cheng and Vassileva 2006, Drenner et al. 2008, Anderson et al. 2013, Burtch et al. 2017, Kokkodis et al. 2019). By combining these results with accurate information about the community welfare from our approach, managers can implement better-informed policy changes. For instance, managers can target promising lurkers (e.g., users in state s_3 of Figure 6) in order to enhance their confidence in the community and accelerate their transition. Alternately, managers can provide incentives (e.g., badges or other community privileges) to engaged users who are fairly likely to lose interest with the community (e.g., users in state s_4 of Figure 6). Importantly, the application of the HMM-AFT framework to six communities shows that the findings are not specific to a single data source.

References

Adaji I, Vassileva J (2016) Toward understanding user participation in stack overflow using profile data. Spiro E, ed. *Internat. Conf. Soc. Informatics* (Springer, Cham, Switzerland), 3–13.

Adler PS, Kwon S-W (2002) Social capital: Prospects for a new concept. *Acad. Management Rev.* 27(1):17–40.

Anderson A, Huttenlocher D, Kleinberg J, Leskovec J (2013) Steering user behavior with badges. *Proc. 22nd Internat. Conf. World Wide Web* (ACM, New York), 95–106.

Angeletou S, Rowe M, Alani H (2011) Modelling and analysis of user behaviour in online communities. *Internat. Semantic Web Conf.* (Springer-Verlag, Berlin, Heidelberg), 35–50.

Arguello J, Butler BS, Joyce E, Kraut R, Ling KS, Rosé C, Wang X (2006) Talk to me: Foundations for successful individual-group interactions in online communities. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (ACM, New York), 959–968.

Bagozzi RP, Dholakia UM (2006) Open source software user communities: A study of participation in linux user groups. *Management Sci.* 52(7):1099–1115.

Bateman PJ, Gray PH, Butler BS (2011) The impact of community commitment on participation in online communities. *Inform. Systems Res.* 22(4):841–854.

Beenen G, Ling K, Wang X, Chang K, Frankowski D, Resnick P, Kraut RE (2006) Using social psychology to motivate contributions to online communities. *Comput.-Mediated Comm.* 10(4):00.

Bernstein MS, Monroy-Hernández A, Harry D, André P, Panovich K, Vargas GG (2011) 4chan and/b: An analysis of anonymity and ephemerality in a large online community. *Proc. 5th Internat. Conf. Weblogs Soc. Media, Barcelona, Catalonia, Spain,* 50–57.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.

Brzozowski MJ, Sandholm T, Hogg T (2009) Effects of feedback and peer pressure on contributions to enterprise social media. *Proc. ACM 2009 Internat. Conf. Supporting Group Work* (ACM, New York), 61–70.

Burke M, Marlow C, Lento T (2009) Feed me: Motivating newcomer contribution in social network sites. *SIGCHI Conf. Human Factors Comput. Systems* (ACM, New York), 945–954.

Burtch G, Hong Y, Bapna R, Griskevicius V (2017) Stimulating online reviews by combining financial incentives and social norms. *Management Sci.* 64(5):2065–2082.

Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* 16(5): 1190–1208.

Cassell J, Huffaker D, Tversky D, Ferriman K (2006) The language of online leadership: Gender and youth engagement on the Internet. *Developmental Psych.* 42(3):436–449.

Chai S, Das S, Rao HR (2011) Factors affecting bloggers' knowledge sharing: An investigation across gender. *J. Management Inform. Systems* 28(3):309–342.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 785–794.

Chen W, Wei X, Zhu K (2018) Engaging voluntary contributions in online communities: A hidden Markov model. *Management Inform. Systems Quart.* 42(1):83–100.

Cheng J, Danescu-Niculescu-Mizil C, Leskovec J (2015) Antisocial behavior in online discussion communities. Preprint, submitted April 2, https://arxiv.org/abs/1504.00680.

Cheng R, Vassileva J (2006) Design and evaluation of an adaptive incentive mechanism for sustained educational online communities. *User Model. User-Adapted Interaction* 16(3–4):321–348.

Cramer M, Hayes G (2010) Acceptable use of technology in schools: Risks, policies, and promises. *IEEE Pervasive Comput.* 9(3):37–44.

Drenner S, Sen S, Terveen L (2008) Crafting the initial user experience to achieve community goals. *Proc. 2008 ACM Conf. Recommender Systems* (ACM, New York), 187–194.

Faraj S, Johnson SL (2011) Network exchange patterns in online communities. *Organ. Sci.* 22(6):1464–1480.

Faraj S, Jarvenpaa SL, Majchrzak A (2011) Knowledge collaboration in online communities. *Organ. Sci.* 22(5):1224–1239.

Faraj S, Kudaravalli S, Wasko M (2015) Leading collaboration in online communities. *Management Inform. Systems Quart.* 39(2): 393–412.

Ferri FJ, Pudil P, Hatef M, Kittler J (1994) Comparative study of techniques for large-scale feature selection. *Machine Intelligence and Pattern Recognition*, vol. 16 (Elsevier, Amsterdam), 403–413.

Forte A, Bruckman A (2005) Why do people write for Wikipedia? Incentives to contribute to open–content publishing. Working paper, Georgia Institute of Technology, Atlanta.

Friedman M, Meiselman D (1963) The relative stability of monetary velocity and the investment multiplier in the United States, 1897–1958. Brown EC, ed. *Stabilization Policies* (Prentice Hall, Englewood Cliffs, NJ), 165–268.

Ghose A, Han S-P (2011) An empirical analysis of user content generation and usage behavior on the mobile Internet. *Management Sci.* 57(9):1671–1691.

Goes PB, Lin M, Au Yeung C-M (2014) "Popularity Effect" in user-generated content: Evidence from online product reviews. *Inform. Systems Res.* 25(2):222–238.

Goldstein JS, Pevehouse JC, Gerner DJ, Telhami S (2001) Reciprocity, triangularity, and cooperation in the middle east, 1979-97. *J. Conflict Resolution* 45(5):594–620.

Grewal R, Lilien GL, Mallapragada G (2006) Location, location, location: How network embeddedness affects project success in open source systems. *Management Sci.* 52(7):1043–1056.

Haines VA, Godley J, Hawe P (2011) Understanding interdisciplinary collaborations as social networks. *Amer. J. Community Psych.* 47(1–2):1–11.

Hay C, Meldrum R, Mann K (2010) Traditional bullying, cyber bullying, and deviance: A general strain theory approach. *J. Contemporary Criminal Justice* 26(2):130–147.

Healey B, Hoek J, Edwards R (2014) Posting behaviour patterns in an online smoking cessation social network: Implications for intervention design and development. *PLoS One* 9(9):e106603.

Hecking T, Chounta I-A, Hoppe HU (2015) Analysis of user roles and the emergence of themes in discussion forums. *2015 2nd Eur. Network Intelligence Conf. (ENIC)* (IEEE, Piscataway, NJ), 114–121.

Hemetsberger A (2002) Fostering cooperation on the Internet: Social exchange processes in innovative virtual consumer communities. *Adv. Consumer Res.* 29:354–356.

Hew KF (2009) Determinants of success for online communities: An analysis of three communities in terms of members' perceived professional development. *Behav. Inform. Tech.* 28(5):433–445.

Himelboim I, Gleave E, Smith M (2009) Discussion catalysts in online political discussions: Content importers and conversation starters. *J. Comput.-Mediated Comm.* 14(4):771–789.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Huang J, Dasgupta A, Ghosh A, Manning J, Sanders M (2014) Superposter behavior in MOOC forums. *Proc. 1st ACM Conf. Learn. Scale Conf.* (ACM, New York), 117–126.

Huang Y, Singh PV, Ghose A (2015) A structural model of employee behavioral dynamics in enterprise social media. *Management Sci.* 61(12):2825–2844.

Huberman BA, Romero DM, Wu F (2009) Crowdsourcing, attention and productivity. *J. Inform. Sci.* 35(6):758–765.

Huffaker D (2010) Dimensions of leadership and social influence in online communities. *Human Comm. Res.* 36(4):593–617.

Hwang EH, Singh PV, Argote L (2015) Knowledge sharing in online communities: Learning to cross geographic and hierarchical boundaries. *Organ. Sci.* 26(6):1593–1611.

Hyndman RJ, Khandakar Y (2007) *Automatic Time Series for Forecasting: The Forecast Package for R* (Monash University, Department of Econometrics and Business Statistics).

Interdonato R, Pulice C, Tagarelli A (2015) Got to have faith!: The devotion algorithm for delurking in social networks. *Proc. 2015 IEEE/ACM Internat. Conf. Adv. Soc. Networks Anal. Mining 2015* (ACM, New York), 314–319.

Interdonato R, Pulice C, Tagarelli A (2016) Community-based delurking in social networks. *Proc. 2016 IEEE/ACM Internat. Conf. Adv. Soc. Networks Anal. Mining* (IEEE, Piscataway, NJ), 263–270.

Isaacs D (2014) Social media and communication. *J. Paediatric Child Health* 50(6):421–422.

Janzik L, Herstatt C (2008) Innovation communities: Motivation and incentives for community members to contribute. *4th IEEE Internat. Conf. Management Innovation Tech. 2008. ICMIT 2008* (IEEE, Piscataway, NJ), 350–355.

Johnson SL, Safadi H, Faraj S (2015) The emergence of online community leadership. *Inform. Systems Res.* 26(1):165–187.

Joyce E, Kraut RE (2006) Predicting continued participation in newsgroups. *J. Comput.-Mediated Comm.* 11(3):723–747.

Kim AJ (2000) *Community Building on the Web: Secret Strategies for Successful Online Communities* (Addison-Wesley Longman Publishing Co., Inc., Boston).

Kohler T, Fueller J, Matzler K, Stieger D (2011) Co-creation in virtual worlds: The design of the user experience. *Management Inform. Systems Quart.* 35(3):773–788.

Kokkodis M (2018) Dynamic recommendations for sequential hiring decisions in online labor markets. *Proc. 24th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 453–461.

Kokkodis M (2019a) Designing dynamic reputation systems for online labor markets. Working paper, Boston College, Boston.

Kokkodis M (2019b) Reputation deflation through dynamic expertise assessment in online labor markets. *World Wide Web Conf.* (ACM, New York), 896–905.

Kokkodis M, Ipeirotis PG (2016) Reputation transferability in online labor markets. *Management Sci.* 62(6):1687–1706.

Kokkodis M, Ransbotham S (2019) Asymmetric reputation spillover from agencies on digital platforms. Working paper, Boston College, Boston.

Kokkodis M, Lappas T, Kane G (2019) Direct and indirect benefits of introducing purchase verification in e-commerce platforms: Evidence from a natural experiment. Working paper, Boston College, Boston.

Koller D, Friedman N (2009) *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA).

Lampe C, Johnston E (2005) Follow the (slash) dot: Effects of feedback on new members in an online community. *Proc. 2005 Internat. ACM SIGGROUP Conf. Supporting Group Work* (ACM, New York), 11–20.

Lampel J, Bhalla A (2007) The role of status seeking in online communities: Giving the gift of experience. *J. Comput.-Mediated Comm.* 12(2):434–455.

Lave J, Wenger E (1991) *Situated Learning: Legitimate Peripheral Participation* (Cambridge University Press, Cambridge, UK).

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. *Proc. Machine Learn. Res* 32(2):1188–1196.

Leimeister JM, Sidiras P, Krcmar H (2006) Exploring success factors of virtual communities: The perspectives of members and operators. *J. Organ. Comput. Electronic Commerce* 16(3-4):279–300.

Lin H-F, Lee G-G (2006) Determinants of success for online communities: An empirical study. *Behav. Inform. Tech.* 25(6):479–488.

Lu Y, Jerath K, Singh PV (2013) The emergence of opinion leaders in a networked online community: A dyadic model with time dynamics and a heuristic for fast estimation. *Management Sci.* 59(8): 1783–1799.

Malinen S (2015) Understanding user participation in online communities: A systematic literature review of empirical studies. *Comput. Human Behav.* 46:228–238.

Mario C, Gould WW, Gutierrez RG, Marchenko Y (2008) *An Introduction to Survival Analysis Using Stata* (StataCorp LP, College Station, TX).

Millington R (2012) *Buzzing Communities: How to Build Bigger, Better, and More Active Online Communities* (FeverBee, London).

Molm LD, Schaefer DR, Collett JL (2007) The value of reciprocity. *Soc. Psych. Quart.* 70(2):199–217.

Moon JY, Sproull LS (2008) The role of feedback in managing the Internet-based volunteer work force. *Inform. Systems Res.* 19(4): 494–515.

Murphy KP (2012) *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA).

Nam KK, Ackerman MS, Adamic LA (2009) Questions in, knowledge in? A study of naver's question answering community. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (ACM, New York), 779–788.

Nonnecke B, Preece J (2001) Why lurkers lurk. *AMCIS 2001 Proc.*, 1521–1530.

Oestreicher-Singer G, Zalmanson L (2013) Content or community? A digital business strategy for content providers in the social age. *Management Inform. Systems Quart.* 37(2):591–616.

Olteanu A, Weber I, Gatica-Perez D (2016) Characterizing the demographics behind the #blacklivesmatter movement. *2016 AAAI Spring Sympos Ser*, 310–313.

Ozturk P, Nickerson J (2015) Paths from talk to action. *36th Internat. Conf. Inform. Systems, Fort Worth, TX,* 1–18.

Pai P, Tsai H-T (2016) Reciprocity norms and information-sharing behavior in online consumption communities: An empirical investigation of antecedents and moderators. *Inform. Management* 53(1):38–52.

Patton DU, Eschmann RD, Butler DA (2013) Internet banging: New trends in social media, gang violence, masculinity and hip hop. *Comput. Human Behav.* 29(5):A54–A59.

Phang CW, Kankanhalli A, Tan BCY (2015) What motivates contributors vs. lurkers? An investigation of online feedback forums. *Inform. Systems Res.* 26(4):773–792.

Preece J (2001) Online communities: Usability, sociability, theory and methods. Earnshaw R, Guedj R, Van Dam A, Vince J, eds. *Frontiers of Human-Centered Computing, Online Communities and Virtual Environments* (Springer-Verlag, London), 263–277.

Preece J, Shneiderman B (2009) The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Trans. Human-Comput. Interaction* 1(1):13–32.

Ransbotham S, Kane GC, Lurie NH (2012) Network characteristics and the value of collaborative user-generated content. *Marketing Sci.* 31(3):387–405.

Ransbotham S, Lurie NH, Liu H (2019) Creation and consumption of mobile word of mouth: How are mobile reviews different? *Marketing Sci.* 38(5):773–792.

Ransbotham S, Fichman RG, Gopal R, Gupta A (2016) Ubiquitous it and digital vulnerabilities. *Inform. Systems Res.* 27(4):834–847.

Ray S, Kim SS, Morris JG (2014) The central role of engagement in online communities. *Inform. Systems Res.* 25(3):528–546.

Reagans R, McEvily B (2003) Network structure and knowledge transfer: The effects of cohesion and range. *Admin. Sci. Quart.* 48(2):240–267.

Ren Y, Harper FM, Drenner S, Terveen LG, Kiesler SB, Riedl J, Kraut RE (2012) Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *Management Inform. Systems Quart.* 36(3):841–864.

Ridings C, Gefen D, Arinze B (2006) Psychological barriers: Lurker and poster motivation and behavior in online communities. *Comm. Assoc. Inform. Systems* 18(1):Article 16.

Saltz JS, Hiltz SR, Turoff M, Passerini K (2007) Increasing participation in distance learning courses. *IEEE Internet Comput.* 11(3): 36–44.

Sapankevych NI, Sankar R (2009) Time series prediction using support vector machines: A survey. *IEEE Comput. Intelligence Magazine* 4(2):24–38.

Schneider A, Von Krogh G, JäGer P (2013) "Whats coming next?" Epistemic curiosity and lurking behavior in online communities. *Comput. Human Behav.* 29(1):293–303.

Schwarz G (1978) Estimating the dimension of a model. *Ann. Statist.* 6(2):461–464.

Seraj M (2012) We create, we connect, we respect, therefore we are: Intellectual, social, and cultural value in online communities. *J. Interactive Marketing* 26(4):209–222.

Shen KN, Khalifa M (2007) Exploring multi-dimensional conceptualization of social presence in the context of online communities. Jacko JA, ed. *Internat. Conf. Human-Comput. Interaction* (Springer, Berlin, Heidelberg), 999–1008.

Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK (2000) Improvements to the SMO algorithm for SVM regression. *IEEE Trans. Neural Networks* 11(5):1188–1193.

Shriver SK, Nair HS, Hofstetter R (2013) Social ties and user-generated content: Evidence from an online social network. *Management Sci.* 59(6):1425–1443.

Sun N, Rau PP-L, Ma L (2014) Understanding lurkers in online communities: A literature review. *Comput. Human Behav.* 38:110–117.

Szmigin I, Canning L, Reppel AE (2005) Online community: Enhancing the relationship marketing concept through customer bonding. *Internat. J. Service Indust. Management* 16(5):480–496.

Tagarelli A, Interdonato R (2013) Who's out there?: Identifying and ranking lurkers in social networks. *Proc. 2013 IEEE/ACM Internat. Conf. Adv. Soc. Networks Anal. Mining* (ACM, New York), 215–222.

Tagarelli A, Interdonato R (2014) Lurking in social networks: Topology-based analysis and ranking methods. *Soc. Network Analysis Mining* 4(1):230.

Tagarelli A, Interdonato R (2015) Time-aware analysis and ranking of lurkers in social networks. *Soc. Network Analysis Mining* 5(1):46.

Tsai H-T, Bagozzi RP (2014) Contribution behavior in virtual communities: Cognitive, emotional, and social influences. *Management Inform. Systems Quart.* 38(1):143–163.

Van Mierlo T (2014) The 1% rule in four digital health social networks: An observational study. *J. Medical Internet Res.* 16(2):e33.

Viégas FB, Smith M (2004) Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. *Proc. 37th Annual Hawaii Internat. Conf. System Sci. 2004* (IEEE, Piscataway, NJ).

Wang Y, Fesenmaier DR (2003) Assessing motivation of contribution in online communities: An empirical investigation of an online travel community. *Electronic Marketing* 13(1):33–45.

Wasko MM, Faraj S (2005) Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *Management Inform. Systems Quart.* 29(1):35–57.

Wiertz C, de Ruyter K (2007) Beyond the call of duty: Why customers contribute to firm-hosted commercial online communities. *Organ. Stud.* 28(3):347–376.

Wikipedia. List of virtual communities with more than 1 million users. Accessed October 24, 2018, https://en.wikipedia.org/wiki/List_of_virtual_communities_with_more_than_1_million_users.

Willard NE (2007) *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress* (Research Press, Champaign, IL).

Wise K, Hamman B, Thorson K (2006) Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. *J. Comput.-Mediated Comm.* 12(1): 24–41.

Wu M (2018) The 90-9-1 rule in reality. Lithium technologies (online). Accessed October 24, 2018, https://lithosphere.lithium.com/t5/Science-of-Social-Blog/The-90-9-1-Rule-in-Reality/ba-p/5463.

XGBoost (2018) Scalable and flexible gradient boosting. Accessed August 24, 2019, https://xgboost.ai/.

Yoo Y, Alavi M (2004) Emergent leadership in virtual teams: What do emergent leaders do? *Inform. Organ.* 14(1):27–58.

Zeng X, Wei L (2013) Social ties and user content generation: Evidence from flickr. *Inform. Systems Res.* 24(1):71–87.

Zhang W, Watts SA (2008) Capitalizing on content: Information adoption in two online communities. *J. Assoc. Inform. Systems* 9(2):Article 3.